ATHENS INFORMATION TECHNOLOGY

CENTER OF EXCELLENCE FOR RESEARCH AND GRADUATE EDUCATION

The RT06 AIT Systems for Speech Activity Detection and Speaker Diarization

E. Rentzeperis, A. Stergiou and A. Pnevmatikakis

{eren, aste, apne}@ait.edu.gr

Athens Information Technology

Autonomic and Grid Computing Group

www.ait.gr

May 4, 2006

Overview

- Available data & usage
- Speech Activity Detection
  - Algorithm
  - Results
  - Comments
- Speaker Diarization
  - Algorithm
  - Results
  - Comments

Evaluation Data

- Two types of meeting recordings
  - Conference meetings
  - Lecture meetings
- Testing data come from various sites with different room and microphone configurations
- Data from previous evaluations were used as
  - training and development for SAD
  - development for SPKR (tune thresholds)

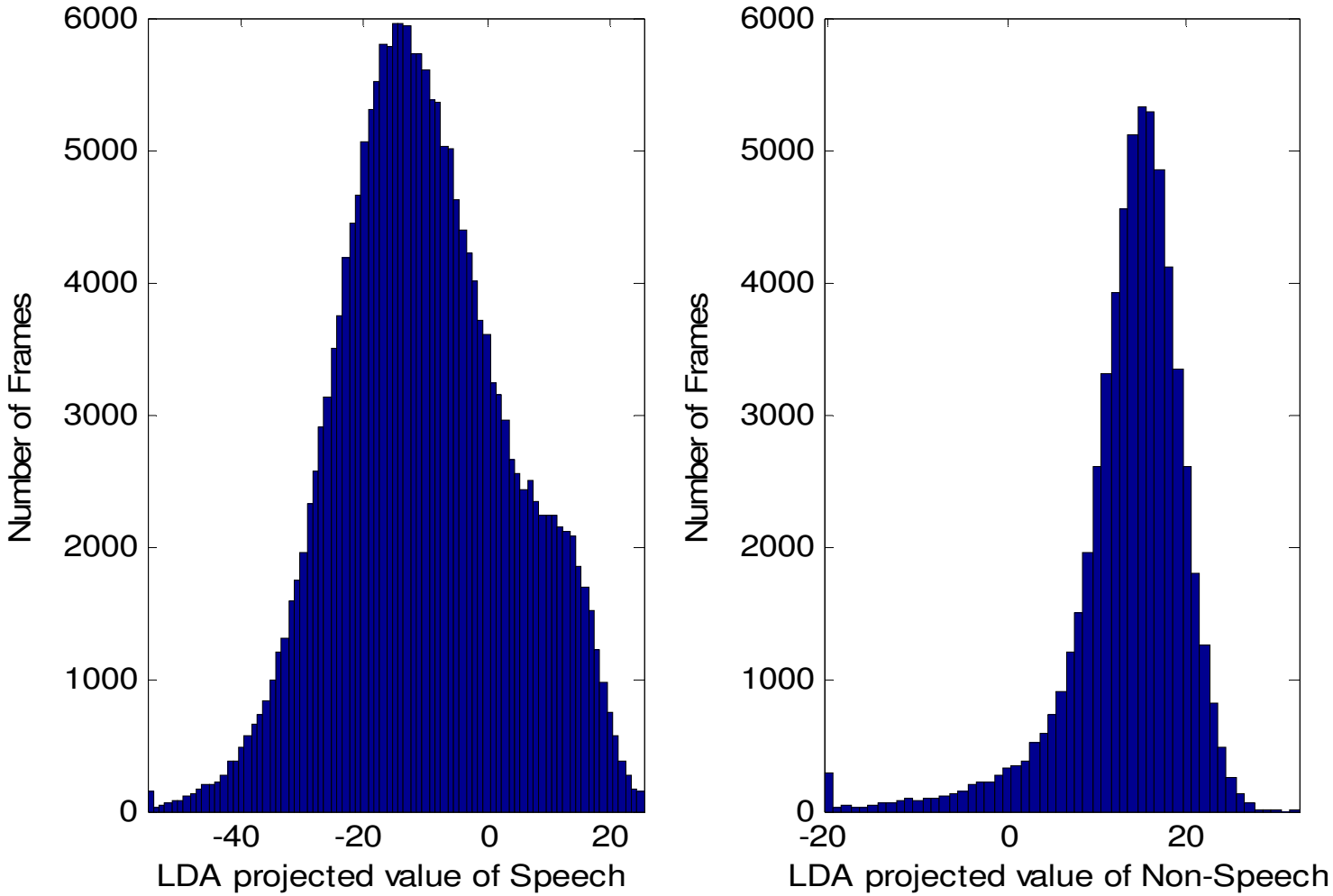
SAD: Introduction

- Purpose of SAD
  - Distinguish between speech and various types of acoustic noise
- Previous Research
  - Energy-based with Adaptive Thresholds
  - Zero-crossing Rate
- Proposed Algorithm
  - Two class (speech / non-speech) Linear Discriminant Analysis applied to Mel Frequency Cepstral Coefficients

SAD: Main Tools

- Mel Frequency Cepstral Coefficients (MFCC)
  - Dominant features used for speech/speaker recognition
  - Obtained by taking the inverse Fourier Transform of the log spectrum after it is wrapped according to a nonlinear Mel scale that is motivated by properties of human hearing
- Linear Discriminant Analysis (LDA)
  - Linear subspace projection aiming at maximizing between class scatter under the constraint of minimum within class scatter
  - Eigenvector corresponding to the non-zero eigenvalue projects 2-class data onto a 1D subspace

SAD: Motivation for the LDA Classifier



SAD: Site-Dependant Training

- The system is allowed to have a priori knowledge of the data collection site, room configuration and sensor types
  - Change training depending on the site and room configuration associated with the data
  - i.e. the training parameters for site Y are evaluated from the development data for Y or from a site that has similar configuration and sensor types to Y
- An ideal scenario would be to have a global training scheme

SAD: Training Procedure

- The individual channel signals are summed up
  - to suppress the silence parts and emphasize the speech intervals
- The sequence is separated to speech and non-speech data according to reference
- The segments are separated into 75% overlapping frames of duration 1024 samples
- The MFCC of each frame are computed
- An eigenvector that maximizes the LDA criterion function is estimated

SAD: Testing Procedure

- Similarly to training
  - The individual channel signals are summed up
  - The spatially averaged audio signal is separated into 75% overlapping frames of duration 1024 samples
  - The MFCC of each frame are computed
- The processed frames are projected in the direction of the LDA eigenvector found in training
- A decision is made based on a heuristic threshold
- Median filtering is applied to smooth decision pattern

SAD: Results for MDM Conditions

- Conference Meetings
  - Overall Error : 10.97%
  - Speed Factor (Total Processing Time/Source Signal Duration) : 0.0095
- Lecture Meetings
  - Overall Error : 13.80%
  - Speed Factor : 0.0478

SAD: Comments

- The system is fast and suitable for real time applications
- The discrimination between speech and non-speech is satisfactory
- However no global training parameters

SAD: Future Directions

- Methods based on LDA
  - Energy-based adaptive method as a preprocessing step
    - Already implemented (MLMI'06)
  - Finite state machine
- Preprocessing that takes advantage of the topology of the microphones
  - Sound localization using information theoretic approaches
  - Beamforming
    - Instead of simply averaging

SPKR: Overview

- System Description for Primary and Contrastive Systems
- Top-Down Segmentation using BIC
- Sequential Bottom-Up Clustering using BIC
- Global Clustering using GMM
- Scores & Processing Time Calculations
- Comments

SPKR: Primary System

- Initial speech segmentation using BIC
  - Many short segments
- Two-step clustering to group segments from the same speaker (hopefully!) together
  - Sequential clustering using BIC
  - Global clustering using GMM
  - Number of speaker clusters is chosen so that they account for more than a heuristically defined percentage of total speech
  - Any remaining segments are assigned to surviving speaker clusters based on the GMM a posteriori log likelihoods

SPKR: Contrastive System

- Initial estimate of speaker change times based on output of the SAD module
- Segmentation using BIC inside the detected speech intervals
- Single-step clustering to group segments from the same speaker together
  - Global clustering using GMM
  - Number of clusters is chosen so that they account for more than a heuristically defined percentage of total speech
  - Any remaining segments are assigned to surviving clusters based on the GMM a posteriori log likelihoods

MFCC

- Input waveform segmented into 1024 sample long frames
  - 75% overlap
- 26-D MFCC extracted
  - 12 static + log-energy + delta coefficients

Segmentation using BIC

- Look for a change in statistics of frames in the current search window
- If no change is detected, increase size of search window by fixed step until a maximum size
  - Search window is reset to minimum, starting point step back to halfway between the minimum and maximum windows, so as not to miss changes near the end of the previous maximum window
- If a change is detected before the maximum window size has been reached
  - Search window is reset to minimum, new starting point is the frame where the change was detected

Sequential Clustering using BIC

- BIC merging of consecutive segments
- Process terminates when either
  - A maximum number of passes has been performed
  - A pass resulted in no merges
- Feeds as large and homogeneous segments as possible to the global GMM clustering step
  - Facilitating GMM training
  - A reduction of about 20-25% in the number of segments is achieved

Global Clustering using GMM

- Segments from previous steps are sorted according to their lengths (descending order)
  - Speaker GMM's are based on as many data as possible
- For each new segment
  - A-posteriori log likelihoods w.r.t. all available GMM's so far are computed
  - If the maximum log likelihood is above a threshold, the segment is assigned to the speaker whose GMM gave that best score
  - Else, this can be a new speaker
    - create a new GMM if the current speaker count is less than a maximum value;
    - else, the segment is assigned to the best matching of the current speakers

Global Clustering using GMM (cont.)

- After all segments have been assigned, the support (number of segments) of each speaker is calculated
- Supports are sorted in descending order and a cumulative support score is computed
- The final number of speakers is such that their cumulative support is more than a heuristically defined threshold

SPKR: Scores & Processing Time

- Overall Speaker Error Rates (%)

confmtg / mdm (p)	confmtg / mdm (c)	confmtg / sdm (p)	lectmtg / mdm (p)	lectmtg / sdm (p)
70.70	66.06	67.22	49.51	45.00

- Processing Time Calculations

Condition	confmtg / mdm (p)	confmtg / mdm (c)	confmtg / sdm (p)	lectmtg / mdm (p)	lectmtg / sdm (p)
TPT (sec)	863.37	930.34	849.52	2344.54	2346.90
SSD (sec)	9723.905	9723.905	9723.905	11400.00	11400.00
SF	0.0888	0.0957	0.0874	0.2057	0.2059

SPKR: Comments

- Significantly better performance for the lectmtg than the confmtg scenario (21.2% for mdm and 22.2% for sdm)
  - Considerably less interruptions and speaker changes due to the nature of the recordings (lectures vs. discussions)
- The use of a SAD module is beneficial to the system (4.64% improvement for the confmtg / mdm condition)
- mdm scores worse than sdm ones (3.5% for confmtg, 4.5% for lectmtg)
  - Simple averaging of all available channels for sdm to suppress noise (spatial averaging)
- System is fast
  - 10-11x real time for confmtg, 5x real time for lectmtg
  - Speed reduced for lectmtg due to higher sampling frequency (44.1 kHz vs. 16 kHz for confmtg)

References

- A. Martin, D. Charlet and L. Mauuary, "Robust Speech/Non-Speech Detection Using LDA Applied to MFCC", ICASSP, 2001
- E. Rentzeperis, et al., "An Adaptive Speech Activity Detector Based on Signal Energy and LDA", MLMI06
- A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian Information Criterion", Eurospeech, pp. 679-682, 1999
- D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, pp. 72 - 83, Jan. 1995